Week10: Exercise

Rajendra Prasad Ponnam

Data Mining, Bellevue University

DSC550-T301: Week10 Assignment

August 13, 2022

Business Problem:

The business problem that the model is built for, is credit risk determination for the loans provided. In a lending organization, in order to mitigate the risk of repayment defaults, the credit worthiness of every loan applicant is determined. The process involves in assessing credit history of the applicant, income, monthly outstanding balance, any delinquency in payment, debt burden etc., In big financial organizations, third party credit score from organizations like Experian, Transunion are used in calculating probability score for the risk. After providing such loans, the loan book is constantly monitored to look for potential charge offs.

Lack of credit risk analysis will lead to organizations providing loans to fraudulent borrowers or borrowers with lack of repayment ability. A charged off loan is a financial loss for the organization. A charged off loan will have to be sold to collections agency and recovery from collections will be significantly lesser than the amount due to be paid.

ML Model benefits:

Organizations program for risk calculations for credit risk management. However, using the machine learning model provides significant benefits over the programmed calculations.

1. Machine learning models can fit even when there is a nonlinear relationship between the predictor variables and the risk.
2. ML models enable the usage of larger number of variables and hence provide better accuracy.
3. A classic programmed calculation will have several assumptions; however, an ML model will not make assumptions and hence provide better insights.
4. ML models have higher predictive power as well as brings in time efficiency for credit decision making.

Data:

Data for model building is obtained from Kaggle. The link to the dataset is provide below. The data dictionary for all the fields are also provided in the link.

https://www.kaggle.com/code/harshitlakhani/credit-risk-eda-on-risky-loans/data?select=loan

The dataset has seventy-four variables and ~887K applications for loans.

The important variables include,

- loan amount
- loan status
- Term
- Revolving Balance amount
- Past delinquency

- Loan grade
- Interest Rate
- Employment length and title
- Purpose
- Home ownership

Project Approach:

I performed the following steps for the model building and evaluation -

1. Performed exploratory data analysis and analyzed the attribute values graphically.
2. Identified correlation between attributes.
3. Identified features required for model building.
4. Identified the best model for the dataset.
5. Built the model and determined evaluation metrics.
6. Derived insights from the model.

Exploratory Data Analysis:

Exploratory data analysis is performed to gain understanding of the dataset. In the loan data set, I performed the following analysis to understand the data.

The data set has loans with multiple loan statuses like charged off, fully paid, current, late by 30 days, at default. The data set is split into multiple data frames by status to perform EDA.

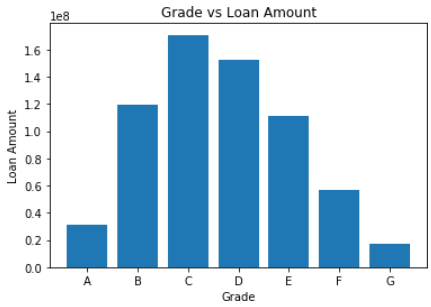
Plotted the following graphs for the applications with both charge off status and fully paid status.

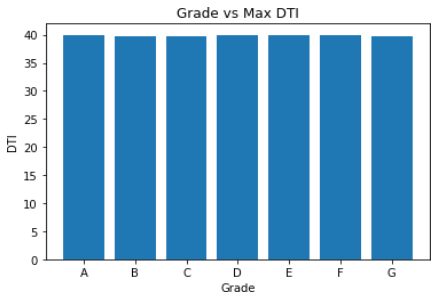
- Grade vs Loan Amount
- Grade vs Max DTI
- Number of loans by term
- Number of loans by verification status
- Number of loans by ownership

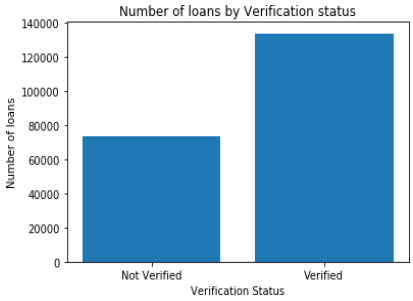
Observations from EDA graphs -

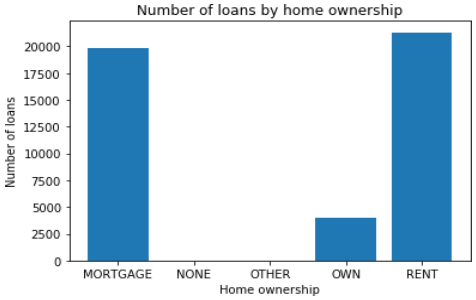
1. All charge off loans have DTI as 40.
2. Fewer A grade loans have charged off as compared to C grade loans.
3. Rate of interest is in proportion to the grade level.
4. In all categories the number of short term loans (36 months) are higher than long term loans.
5. Number of C grade loans are high in current status and have higher probability to default.

The EDA graphs look for charged off loans look like,

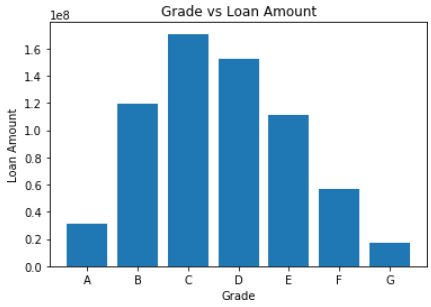


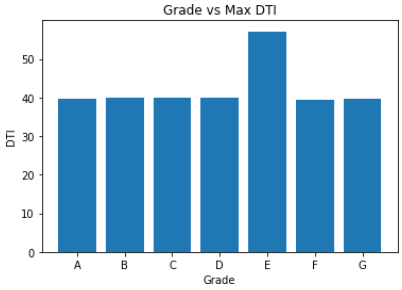


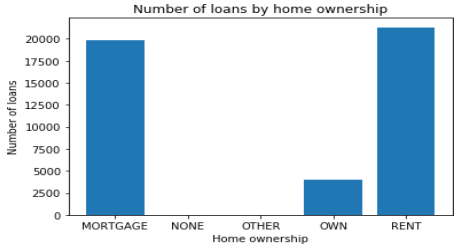


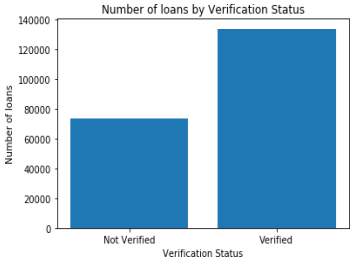


The EDA graphs look for fully paid loans look like,









Data Transformations:

In order to model the data, I performed the following data transformations in the data.

1. Split the data frame into multiple data frames based on status.
2. Summarized the loan amount by grade
3. Conformed the field verification status by replacing the value “Source Verified” as “Verified”.
4. Dropped the fields with null values for more than 97% of records.

Feature Selection:

As a part of feature selection, first converted the categorical features in dummies using pandas get_dummies function. Setting the loan_status_charged_off as the target variable, scaled the numeric features using standardscalar. Using SequentialFeatureSelector function, seven most important features among numerical features are identified. Selected top ten categorical features using chi2.

Selected Numerical features include, funded_amnt, delinq_2yrs, total_pymnt, total_rec_prncp, total_rec_int, recoveries, mths_since_last_major_derog

Selected categorical features include, term_ 60 months, grade_B, grade_C, grade_D, grade_E, grade_F, grade_G, verification_status_Verified, home_ownership_MORTGAGE, home_ownership_RENT.

Model Selection and Evaluation:

As the objective of the project is whether a loan provided will likely be charged off or not, is a classification problem, used a classification model. Using sklearn's traintestsplit, the dataset is split into Xtrain, Xtest, ytrain and ytest. GridSearchCv iwas used to determine the model that will provide more accurate results among - Logistic Regression and Random Forest.

GridSearchCV passed the training dataset through both logistic regression and random forest models and determined that logistic regression is the best model for determining if a loan will likely be charged off. The logistic regression model also has the best model score of 82.22%

However, while evaluating confusion matrix metrics, it is observed that the precision for logistic regression is just 1, while the same for Random forest is 40. The accuracy of Random forest is 81.88%, not much lesser than Logistic Regression. Hence, Random forest seems to be a better model for the objective with the chosen dataset.

Evaluation metrics

The evaluation metrics for the model is as below,

```
Accuracy: 0.8187970915537667
Precision: 0.4
Recall: 0.012773601224136785
```

Changes as compared to prior milestone:

The recall score for the model is too low, indicating imbalanced classes. Charge off is the majority class and the paid off loans are the minority class. In order to balance the classes, utilized SMOTE.

After balancing the dataset using SMOTE, reran randomforest classifier model. The confusion matrix and the evaluation metrics look much improved.

Confusion Matrix:

```
[[68299,   162],
       [  103, 68534]])
```

Evaluation metrics:

```
Accuracy: 0.9980670761061431
Precision: 0.9976417840922325
Recall: 0.9984993516616403
```

Challenges:

The biggest challenge with the project is in the dataset, that contained significantly higher number of records for charge off status as compared to other statuses. Hence, the training the other class was not adequate.

Conclusion:

Using randomforestclassifier model built for the dataset, for any new loan application or an existing issued loan, the possibility of charge off can be determined. If a new loan application has the possibility of charge off, then the application can be rejected. If a current loan has the possibility of charge off, the customer can be approached to provide additional flexibility to avoid the charge off. Thus, the model when deployed in production, can effectively help in credit risk management.